DOCUMENT RESUME

ED 398 271 TM 025 403

AUTHOR Lam, Peter; Foong, Yoke-Yeen
TITLE Rasch Analysis of Math SOLO Taxonomy Levels Using Hierarchical Items in Testlets.
PUB DATE [96]
NOTE 19p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classification; Computer Assisted Testing; Foreign Countries; Goodness of Fit; Item Banks; *Item Response Theory; *Mathematics; Psychometrics; Rating Scales; Secondary Education; *Secondary School Students; Test Construction; *Test Interpretation
IDENTIFIERS Partial Credit Model; *Rasch Model; Rating Scale Model; *SOLO Taxonomy; Testlets

ABSTRACT

This study attempts to estimate Structure of Learning Outcome (SOLO) levels in mathematics using the Partial Credit and Rating Scale models. A 30-item test comprising 10 testlets of 3 items each was designed and administered to 674 lower secondary school students. The items were arranged in a hierarchical manner, each testing SOLO levels in this order: Unistructural, Multistructural, and Relationship/Abstract. The item response matrix was fitted into the Partial Credit and Rating Scale models. Results showed that: (1) the observed testlet response patterns fitted those expected; (2) the dataset fitted the psychometric models reasonably well; and (3) the proportion of examinees getting SOLO items correct decreased in order from level 1 to level 3 along the math proficiency continuum between -2.0 and +2.0. The results of the study have implications for a criterion-based approach in interpreting test results based on SOLO testlets. Results also showed the viability of testlet item bank development, test construction, and computerized testing using testlets. (Contains 1 figure, 4 tables, and 20 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 398 271

RASCH ANALYSIS OF MATH SOLO TAXONOMY LEVELS USING HIERARCHICAL ITEMS IN TESTLETS

Main Author : Peter Lam
National Institute of Education
Center for Educational Research

469 Bukit Timah Road
S259756 SINGAPORE

Coauthor : Yoke-Yeen Foong
School of Science
National Institute of Education

469 Bukit Timah Road
S259756 SINGAPORE

Suggested running head: SOLO estimation using testlets and IRT

Key words: Testlets, IRT, Partial Credit Model, SOLO Taxonomy

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

PETER LAM

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

TM025403

BEST COPY AVAILABLE

RASCH ANALYSIS OF MATH SOLO TAXONOMY LEVELS USING HIERARCHICAL ITEMS IN TESTLETS

ABSTRACT

This study attempts to estimate SOLO (Structure of Learning Outcome) levels in math using the Partial Credit and the Rating Scale models. A 30 item test comprising 10 testlets of 3 items each was designed and administered to 674 lower secondary school students. The items were arranged in a hierarchical manner, each testing SOLO levels in the order: Unistructural, Multistructural and Relational/Abstract. The item response matrix was fitted into the Partial Credit and Rating Scale models. Results showed that a) The observed testlet response patterns fitted those of the expected, b) The dataset fitted the psychometric models reasonably well and c) The proportion of examinees getting SOLO items correct decreased in the order: Level 1, Level 2 and Level 3 along the math proficiency continuum between -2.0 and +2.0. The results of the study have implications for a criterion-referenced approach in interpreting test results based on SOLO testlets. The results also showed the viability of testlet item bank development, test construction and computerized testing using testlets.

RASCH ANALYSIS OF MATH SOLO TAXONOMY LEVELS USING HIERARCHICAL ITEMS IN TESTLETS

INTRODUCTION

It has been argued that current external assessments for accountability may be taking precedence over other educational assessment purposes leading to among others, a de-emphasis on higher order thinking skills (Suarez & Gottovi, 1992). The move towards authentic assessment is an attempt to address this issue. One form of assessment involves methods that emphasize strategies that allow students to plan, monitor, and evaluate their own learning (Burke, 1993; Kiernan & Pyne, 1993). Andrada and Linden (1993) in their study of the psychometric properties of objective tests designed to measure three levels of Bloom's Taxonomy (1956), found that well developed objective tests measuring higher-order thinking can function effectively even when students are allowed to sit for the test under take-home conditions. The study found that the psychometric properties of a test measuring higher order skills can remain intact even when students have ample time and course materials available.

A recent measurement technique involves the use of item clusters or testlets that tap higher level thinking. This kind of test format is not new as it was originally designed to address issues such as context effects, item ordering and content-balancing with practical implications for computerized and adaptive testing (Wainer & Kiely, 1987). Recent development of reading assessment involves the construction of testlets, each associated with a textual passage. The items in the testlets are designed to test various higher order thinking skill levels and are therefore of different difficulty levels

(Crehan *et al.*, 1993). The Illinois Goal Assessment Program (IGAP) reading test uses textual passages and 15 testlets, each containing 5 items, associated with each passage (Wang & Ackerman, 1994). Each testlet requires students to demonstrate various levels of cognitive skills. A feature of this test is that it uses a multiple-response (or multiple-correct) rather than multiple-choice format.

On the psychometric perspective, this form of testing leads to inter-item dependence within testlets and inflated reliability estimates when items are treated as stand-alone or unrelated. By using testing units that are larger than the test item, item dependency can be reduced although the use of testlets does not prevent dependency between testlets (Wainer & Lewis, 1990). In the case of this study where the 30-item math test consists of testlets, each with hierarchical items testing increasing levels of thinking, and each associated with a common stimulus material or mathematical concept to be tested, there is the possibility of violation of unidimensionality of the response matrix as well as item independence if the response matrix is treated as a set of responses to 30 individual items instead of 10 testlets. This loss of unidimensionality has been demonstrated by Thissen, Steinberg and Mooney (1989) using the response matrix of 22 individual reading comprehension items which was originally grouped into 4 testlets, each associated with a reading passage. Analysis using full-information factor analysis resulted in four factors.

Thissen, Steinberg and Mooney (1989) proposed and successfully tested Bock's (1972) polychotomous response model on testlets. Since then, a number of studies have been carried

out on testlets, including testlet validity based scoring (Wainer *et al.*, 1992), testlet-based computerized mastery testing (Sheehan & Lewis, 1992) and testlet reliability (Sireci *et al.*, 1991). While the results of these studies are very appealing to the measurement specialist, there is a need to make IRT and testlets more meaningful, especially to the practitioner interested in assessing higher order thinking skills. As such, earlier studies involving hierarchical items within testlets can be further explored in the context of problem solving skills and learning outcomes within the IRT framework.

This study attempts to apply IRT principles in assessing learning outcome proficiency in math. It is based on Wainer & Kiely's (1987) definition of testlet as a group of items related to a single content area and consisting of a set of predetermined response paths. The same term in this study is defined as a group of 3 hierarchical items testing three SOLO (Structure of Learning Outcome)(Biggs & Collis, 1982) levels. The study is part of a broader study to foster the application of IRT in schools and to make analysis using IRT more meaningful to the class teacher who is well proficient in the application of SOLO Taxonomy in math testing.

METHOD

An example of a SOLO testlet designed for this study is as follows:

> The computer program is able to multiply any number by four and then add three to the result. For example, if you key in the number, '3' the answer is '15'.

Level 1: Unistructural

If you key in the number '5' what is the answer?

A)3 B)12 *C)15 D)18

Level 2: Multistructural

If the computer gives '27' as the answer, what was the number that you keyed in?

*A)6 B)24 C)27 D)111

Level 3: Relational or Abstract

If x is the number that you keyed in and y is the answer that the computer gives out, write a formula that will give us the value of y whatever the value of x.

A) 4y + x = z
*B) 4x + 3 = y
C) 4x + y = z
D) 3x + 4 = y

Each SOLO testlet consists therefore of three items arranged hierarchically, each item testing a specific SOLO level.

As in the case of comprehension passages, the common stimulus material may result in a loss of local independence and hence, possibly the loss of unidimensionality of the test. By grouping hierarchical items within testlets, IRT can be applied to estimate thresholds pertaining to SOLO attainment levels.

A 10-testlet instrument on word problems in math comprising 3 items in each testlet was administered to 674 Lower Secondary pupils. The 30 items were scored individually and the examinee responses (1 = correct, 0 = wrong) submitted to a binary factor analysis using the computer program, MicroFACT (Waller, 1995). Item scores were linearly summed to obtain the testlet scores. Based on the polychoric correlations, the testlet dataset was factor analyzed using the same program to determine its dimensionality. The dataset was fitted into Masters' (1982) Partial Credit and Andrich's (1978) rating scale models using

the computer program, QUEST (Adams & Khoo, 1992). The Quest program uses a generalised form of Masters' (1982) Partial Credit Model (Wright & Masters, 1982) which takes into account, null categories in the estimation process:

$$P(X_{ni}=x_{ni}) = \frac{\exp\sum_{j=0}^{x_{ni}} w_{ij}(\beta_n-\delta_i-\tau_{ij})}{\sum_{k=0}^{mi}\exp\sum_{j=0}^{k} w_{ij}(\beta_n-\delta_i-\tau_{ij})}$$

where β_n is the proficiency level of person n,

w_{ij} is the score assigned to step j in item i to allow for estimation where null categories are involved, and

δ_i and τ_{ij} characterise the difficulty or location parameter of item i.

The following constraints are applied:

$$\exp\sum_{j=0}^{0} w_{ij}(\beta_n-\delta_i-\tau_{ij})=1; \quad \sum_{j=0}^{m_i}\tau_{ij}=0; \quad \sum_{i=1}^{I}\delta_i=0$$

If the τ parameters are constrained so that $\tau_{1j} = \tau_{2j} = \tau_{3j} = \ldots$, then the Partial Credit Model becomes the Andrich Rating Scale Model.

RESULTS

Table 1 shows the response patterns of the 10 testlets. For each testlet with 3 hierarchical items, the expected scoring patterns of the testlets are:

000, 100, 110, 111

The first pattern results from examinees getting all three items in the testlet wrong. The second pattern results from examinees getting the easiest item (i.e. lowest SOLO level) correct and the next two harder items wrong. The third pattern

results from examinees getting the items in the first two SOLO levels correct and the highest SOLO level wrong. The last pattern results from examinees getting all three items in the testlet correct. Departures from the expected pattern (001, 010, 101, 011) are possible due to guessing or other random responses.

Table 1 shows the response patterns and the percentage of examinees for all 10 testlets. The table shows the significantly higher percentages of the expected patterns based on the hierarchical arrangement of the items testing their respective SOLO levels.

Insert Table 1 about here

Table 2 shows the classical item proportion correct (p) and item discrimination (d) values. Within each testlet, the p values showed a general trend of increasing difficulty levels of the hierarchical items. Information from both tables indicates that the responses of the examinees to the items met the expectations of the test constructors in the design of hierarchical items testing different levels of the SOLO Taxonomy.

Insert Table 2 about here

Results of the principal factor analysis using the tetrachoric correlations showed a two-factor solution with a minor second factor in which 8 items loaded significantly on this factor (see Table 3). Items 9, 12, 21, 27 and 30 which loaded significantly on the second factor were the third items of Testlets 3, 4, 7, 9 and 10 respectively. These items test the

SOLO Level 3 (relational/abstract) level. The lower proportion correct values of these items, (p = 0.25, 0.17, 0.19, 0.29 and 0.22) were an indication of their higher difficulty levels. The remaining 3 items (8, 11 and 26) were the second items of Testlets 3, 4 and 9 which were the same 3 testlets identified earlier as having their third items loading the second factor. This second factor which can be defined as the 'difficulty factor' is identified as the 'extended abstract factor' based on Biggs and Collis' (1982) highest SOLO level. Most other items had very minor loadings or almost no loading on the second factor. Item 6 did not load on either factors probably due to its very low p and d values. The item tests on the ability to formulate a scientific equation that will explain the observed phenomenon.

The presence of this difficulty factor indicates that the item response matrix is not unidimensional.

Insert Table 3 about here

The item response matrix was converted to the testlet response matrix by summing the item scores in each testlet. Factor analysis based on the tetrachoric correlations yielded a one-factor solution. When a two factor solution was invoked, the variance explained by the first factor was 5.07 and the second was 0.31 which was not significant. Based on the results of the testlet loadings and the eigenvalue plot, testlet response matrix was essentially unidimensional.

The dataset was then submitted to a Partial Credit and Rating Scale analyses. Analysis of the infit (weighted) and outfit (unweighted) statistics showed that the expected value of

the mean squares was close to 1 and the expected value of the mean t statistics was close to 0 (see Table 4). The infit mean square for each item fell within the acceptable range of 30% above and 30% below the expected value of 1.0. The reliability of estimate was 0.86. This is the proportion of the observed estimate variance that is considered true. Based on these results, the dataset fitted both models reasonably well.

Insert Table 4 about here

Table 4 also shows the estimated SOLO thresholds of the Partial Credit and Rating Scale models. Because of the mathematical relationship between the two models, the estimated SOLO item threshold differences (i.e. Level 2 - Level 1, Level 3 - Level 2) are the same for both models. However, item thresholds will differ from one model to the other for each respective item. The difficulty of crossing the unistructural threshold depends on the complexity of the word problems of the first items in each testlet. Testlets 3, 4, 9 and 10 have their first items testing very basic arithmetic concepts and tend to be easy compared to their respective second items which test the multistructural level. Testlets 2, 3, 4 and 7 have vast differences between the second and third items. Hence, crossing the unistructural SOLO level towards the multistructural SOLO level would require a higher overall math proficiency for examinees attempting Testlets 3, 4, 9 and 10. By the same token, crossing the multistructural SOLO level towards the abstract level requires a much higher level of proficiency for examinees attempting Testlet 2, 3, 4 and 7. Testlets in which the 3 component items are very close in difficulty levels are those of

Testlets 5 and 8. Testlets in which threshold differences are small between the first and the second SOLO levels as measured by the items but large between the second and third levels are those of Testlets 2 and 7. On the other hand, testlets in which the threshold differences are large between the first and second SOLO levels but small between the second and third levels are those of 1 and 9.

Figure 1 shows a plot of SOLO scores with proficiency estimates. The plot shows the presence of differences in item scores between SOLO levels along the math proficiency continuum. This difference increases as proficiency level increases from -2.0 to 2.0 and decreases for math proficiency less than 2.0 and more than 2.0. The plot shows the relative difficulties of the 3 SOLO levels. Judging by the steepness of the plots, as proficiency increases, the rate of getting items at the SOLO Level 1 correct is greater than those in SOLO Level 2 between $-2.0 < \theta < 2.0$. In like manner, the rate of getting SOLO Level 2 items correct is greater than those in SOLO Level 3 as proficiency increases. The small differences beyond +2.0 or -2.0 are probably due to floor and ceiling effects as a result of the responses of very high or very low ability level examinees.

Insert Figure 1 about here

DISCUSSION AND CONCLUSION

The results of the study demonstrated the application of IRT within the framework of pedagogy and testing by the use of hierarchical items in testlets. An important pedagogical consideration is the assessment of learning objectives. The

study showed that as proficiency increases, the number of items answered correctly in each SOLO level increases along the math proficiency continuum between -2.0 and +2.0. However, the number of SOLO Level 1 items answered correctly in SOLO Level 1 is always higher than that of Level 2 and that of Level 2 is always higher than that of Level 3. The practitioner can use a criterion-referenced approach to interpreting test results. For example, if the criterion for attainment of the Unistructural Level in problem solving is at least a minimum of 7 SOLO Level 1 items based on the teacher's judgement, examinees are likely to have an overall math proficiency level of at least -0.4. However, these examinees are likely to have at most 5 SOLO Level 2 items correct or at most 3 SOLO Level 3 items or less correct. To attain the minimum of 7 SOLO Level 3 items correct, the examinees' overall math proficiency must be above 2.0. This would mean that the examinees would score almost all SOLO Level 1 and 2 items correct. This criterion-referenced approach helps the teacher to make instructional decisions such as identifying students for remedial or supplementary work in enhancing problem solving skills. For example, for students with math proficiency < 1.0, the teacher may well emphasize more the unistructural problem solving skills first. Students with math proficiency > 2.0 may well be given most problems involving the extended abstract level.

The results of the study also have implications for the use of testlets in item banking and test construction. It can be seen that the threshold levels vary widely among the 10 testlets. Given a large set of calibrated testlets to form a testlet bank, individual testlets can be selected to define a

suitable range of threshold levels. For example, if the test constructor decides that the thresholds for the unistructural, multistructural and relational/abstract levels should be in the region of -1.0, 0 and 2.5 logits respectively, then items 2 and 7 would be most suitable. In this study, it is expected that given the small number of testlets, the choices would be very limited. Given a large testlet bank appropriately calibrated, it is possible to develop tests adaptively from testlets, each estimating three SOLO Taxonomy levels leading to an extended version of adaptive tests that integrate cognitive performance with IRT.

Table 1. Response Patterns and Percentage Responding for 10 Testlets

Pattern	1	2	3	4	5	6	7	8	9	10
000	8.3	22.4	29.5	16.5	11.6	10.5	19.3	12.3	16.3	14.1
100	10.2	12.9	23.4	36.1	4.0	11.0	13.5	7.3	31.0	26.1
110	15.0	36.7	13.5	28.0	12.2	25.7	35.8	15.0	17.7	31.5
111	52.5	7.7	5.8	9.6	62.9	42.9	8.8	52.8	11.6	15.3
	85.7	79.7	72.2	90.2	90.7	90.1	77.4	87.4	76.6	87.0
001	2.7	6.1	6.4	1.9	2.2	2.3	4.3	3.7	5.2	1.5
010	2.8	8.7	8.3	2.1	0.9	3.0	12.0	2.4	6.2	6.6
101	7.9	2.1	9.8	4.6	3.7	2.7	2.9	4.6	10.7	3.7
011	0.9	3.4	3.3	1.2	2.5	1.9	3.4	1.9	1.3	1.2
	14.3	20.3	27.8	9.8	9.3	9.9	22.6	12.6	23.4	13.0

N = 674

Table 2. Classical Item p and d Values

Item #	Testlet #	p	d	Item #	Testlet #	p	d
1	1	0.85	0.52	16	6	0.82	0.49
2	1	0.71	0.52	17	6	0.73	0.62
3	1	0.64	0.56	18	6	0.50	0.54
4	2	0.59	0.62	19	7	0.61	0.62
5	2	0.57	0.57	20	7	0.60	0.56
6	2	0.19	0.10	21	7	0.19	0.16
7	3	0.53	0.54	22	8	0.80	0.64
8	3	0.31	0.33	23	8	0.72	0.65
9	3	0.25	0.30	24	8	0.63	0.59
10	4	0.78	0.62	25	9	0.71	0.68
11	4	0.41	0.45	26	9	0.37	0.37
12	4	0.17	0.22	27	9	0.29	0.30
13	5	0.83	0.59	28	10	0.77	0.62
14	5	0.79	0.63	29	10	0.55	0.43
15	5	0.71	0.51	30	10	0.22	0.32

N = 674

Table 3. Sorted Factor Loadings (Promax Rotated) based on Tetrachoric Correlation Matrix of the 30 Items

	Factor 1	Factor 2		Factor 1	Factor 2
I22	-0.94	-0.16	I3	-0.60	0.12
I23	-0.87	-0.09	I20	-0.55	0.18
I17	-0.83	-0.08	I5	-0.54	0.25
I25	-0.82	0.03	I18	-0.53	0.20
I13	-0.82	-0.02	I15	-0.53	0.16
I14	-0.80	0.03	I29	-0.33	0.26
I28	-0.80	0.00	I6	-0.08	0.03+
I10	-0.73	0.13	I27	-0.08	0.63*
I1	-0.72	0.02	I12	0.14	0.61*
I16	-0.72	-0.09	I30	-0.04	0.59*
I24	-0.72	0.01	I11	-0.25	0.45*
I4	-0.72	0.09	I8	-0.11	0.40*
I19	-0.70	0.10	I26	-0.24	0.31*
I2	-0.65	-0.01	I9	-0.19	0.28*
I7	-0.62	0.06	I21	-0.01	0.22*

Variance explained Factor 1: 10.10 Factor 2: 3.25

* Significant loadings on second factor
+ Nonsignificant loadings on both factors

Table 4. Partial Credit and Rating Scale Analysis of Testlet Dataset

Testlet		Thresholds 1		2		3		Inft MNSQ	Inft t	Outft MNSQ	Outft t
1	PC	-1.87		-0.82		-0.27		1.03	0.50	1.21	2.00
	RS	-0.88		0.17		0.71					
	Diff		1.05		0.55						
2	PC	-0.80		-0.03		2.90		1.00	0.10	1.03	0.50
	RS	-1.49		-0.72		2.21					
	Diff		0.77		2.93						
3	PC	-0.58		0.97		3.00		1.06	1.10	1.09	1.30
	RS	-1.71		-0.16		1.87					
	Diff		1.55		2.03						
4	PC	-1.46		0.51		2.51		1.07	1.40	1.07	1.00
	RS	-1.98		-0.01		1.99					
	Diff		1.97		2.00						
5	PC	-0.85		-1.42		-0.83		1.03	0.50	1.09	0.70
	RS	0.19		-0.38		0.20					
	Diff		0.57		0.59						
6	PC	-1.55		-0.87		0.27		0.97	-0.50	0.97	-0.40
	RS	-0.83		-0.15		0.99					
	Diff		0.68		1.14						
7	PC	-1.06		-0.01		2.75		0.94	-1.30	0.93	-1.00
	RS	-1.62		-0.57		2.19					
	Diff		1.05		2.76						
8	PC	-1.22		-0.82		-0.35		0.78	-3.90	0.79	-2.00
	RS	-0.42		-0.02		0.44					
	Diff		0.40		0.47						
9	PC	-1.51		0.69		2.20		0.91	-1.80	0.91	-1.40
	RS	-1.97		0.23		1.74					
	Diff		2.20		1.51						
10	PC	-1.59		0.14		1.96		1.07	1.30	1.07	1.10
	RS	-1.76		-0.03		1.79					
Threshold	Diff		1.73		1.82		Mean	0.97	-0.06	1.02	0.11

PC = Partial Credit
RS = Rating Scale

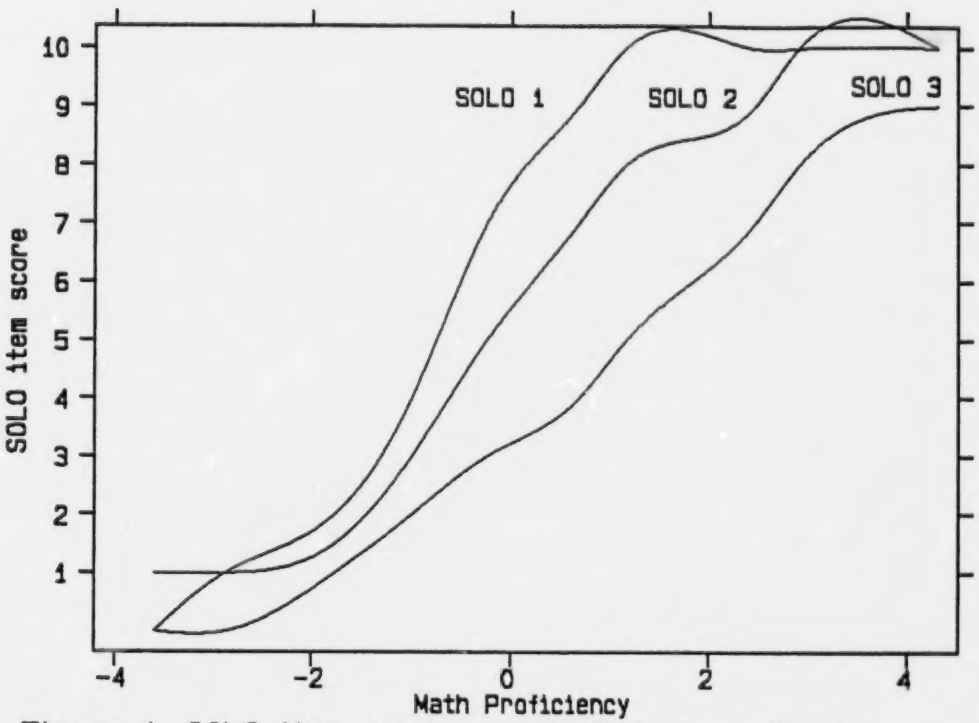

Figure 1. SOLO item scores vs proficiency estimates

BIBLIOGRAPHY

Adams, R.J. & Khoo, S.T. (1994). QUEST Manual. Vic.: Australian Council of Educational Research.

Andrada, G.N. & Linden, K.W. (1993). Effects of two testing conditions on classroom achievement: Traditional in-class versus take-home conditions. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.

Andrich, D.A. (1978). A rating formulation for ordered response categories. Psychometrika,43,561-573.

Biggs, J.B. & Collis, K.F. (1982). Evaluating the Quality of Learning. New York: Academic Press.

Bloom, B.S. (Ed.) (1956). Taxonomy of Educational Objectives: Handbook I, Cognitive Domain. New York: David Mckay.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37,29-51.

Burke, K. (1993). The Mindful School: How to Assess Thoughtful Outcomes. Ill:IRI/Skylight Publishing.

Crehan, K.D. & others (1993). A comparison of testlet reliability for polychotomous scoring methods. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta.

Kiernan, H. & Pyne, J. (1993). Measuring student performance: Assessment in the Social Studies. The Docket: Journal of the New Jersey Council for the Social Studies, Winter 1993.

Masters, G.N. (1982). A Rasch model for partial credit scoring. Psychometrika,47,149-174.

Sheehan, K. & Lewis, C. (1992). Computerized mastery testing with nonequivalent testlets. Applied Psychological Measurement,16,65-76.

Sireci, S.G. *et al.* (1991). On the reliability of testlet-based tests. Journal of Educational Measurement,28,23-47.

Suarez, T.M. & Gottovi, N.C. (1992). The impact of high-stake assessments on our schools. NASSP-Bulletin,76,82-88.

Thissen, D., Steinberg, L. & Mooney, J.A. (1989). Trace lines for testlets: A use of the multiple-categorical-response models. Journal of Educational Measurement,26,247-260.

Wainer, H. & Kiely, G.L. (1987). Item clusters and computerized adaptive testing. Journal of Educational Measurement,24, 185-201.

Wainer, H. & Lewis, C. (1990). Toward a psychometrics for testlets. Journal of Educational Measurement, 27, 1-14.

Wainer, H. *et al.* (1990). A testlet-based, hierarchically-structured test with validity-based scoring. Technical Report No. 90-92. Educational Testing Service. N.J.: Educational Testing Service.

Waller, N.G. (1995). MicroFACT Manual. St Paul: Assessment Systems Corporation.

Wang, C.S. & Ackerman, T. (1994). An examination of response dependency when there is more than one correct answer. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Wright, B.D. & Masters, G.N. (1982). Rating Scale Analysis: Rasch Measurement. Chicago: MESA Press.